Title
EcoCyc: fusing model organism databases with systems biology.

Journal
Nucleic acids research, 41(Database issue)

ISSN
0305-1048

Authors
Keseler, Ingrid M
Mackie, Amanda
Peralta-Gil, Martin
et al.

Publication Date
2013

DOI
10.1093/nar/gks1027

Peer reviewed

Published online 9 November 2012 *Nucleic Acids Research, 2013, Vol. 41, Database issue* ***D605–D612***
doi:10.1093/nar/gks1027

EcoCyc: fusing model organism databases with systems biology

Ingrid M. Keseler[1,*], Amanda Mackie[2], Martin Peralta-Gil[3], Alberto Santos-Zavaleta[3], Socorro Gama-Castro[3], César Bonavides-Martínez[3], Carol Fulcher[1], Araceli M. Huerta[3], Anamika Kothari[1], Markus Krummenacker[1], Mario Latendresse[1], Luis Muñiz-Rascado[3], Quang Ong[1], Suzanne Paley[1], Imke Schröder[4,5], Alexander G. Shearer[1], Pallavi Subhraveti[1], Mike Travers[1], Deepika Weerasinghe[1], Verena Weiss[3], Julio Collado-Vides[3], Robert P. Gunsalus[4,5], Ian Paulsen[2] and Peter D. Karp[1,*]

[1]SRI International, 333 Ravenswood Avenue, Menlo Park, CA 94025, USA, [2]Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney, New South Wales 2109, Australia, [3]Programa de Genómica Computacional, Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, A.P. 565-A, Cuernavaca, Morelos 62100, México, [4]Department of Microbiology, Immunology, and Molecular Genetics and [5]UCLA Institute of Genomics and Proteomics, University of California, Los Angeles, CA 90095, USA

Received September 15, 2012; Accepted October 4, 2012

ABSTRACT

EcoCyc (http://EcoCyc.org) is a model organism database built on the genome sequence of *Escherichia coli* K-12 MG1655. Expert manual curation of the functions of individual *E. coli* gene products in EcoCyc has been based on information found in the experimental literature for *E. coli* K-12-derived strains. Updates to EcoCyc content continue to improve the comprehensive picture of *E. coli* biology. The utility of EcoCyc is enhanced by new tools available on the EcoCyc web site, and the development of EcoCyc as a teaching tool is increasing the impact of the knowledge collected in EcoCyc.

OVERVIEW

EcoCyc has a long history of capturing *Escherichia coli* biology. In 1994, expert manual curation of EcoCyc began by covering the area of metabolic pathways and enzymes. Since then, EcoCyc has evolved in both breadth and depth: it now incorporates the functional annotation of all gene products, including proteins and RNAs outside of metabolic pathways. Many new data types, such as evidence codes, signaling pathways, transcriptional and post-transcriptional regulation and Gene Ontology (GO) annotations, have been added by curators. Highlights of our progress in updating EcoCyc content and the functions of *E. coli* gene products are described later and summarized in Table 1.

We propose that a next step in the evolution of model organism databases (MODs), such as EcoCyc, is to become computational models of their respective organisms. We have generated a steady-state metabolic flux model from EcoCyc using the MetaFlux (1) implementation of flux balance analysis (FBA), and we have used that model to predict the growth phenotype (growth or no growth) of *E. coli* under many different nutrient and gene knockout conditions. We are undertaking a long-term iterative effort to perform these computational predictions, to compare the computational results to experimental results, and to investigate the differences between the two. We will update the metabolic reaction model within EcoCyc when warranted to resolve these differences, and we hope that our efforts will lead to new experimental investigations in cases where the phenotypic observations cannot be explained.

One merit of our proposed marriage of MODs with systems biology is to yield higher-quality databases—in fact, it has already led to improvements in EcoCyc, such as the addition of previously overlooked reactions from the literature and the correction of reaction direction information. Subjecting a database to computational consistency checks can identify errors that manual analysis overlooks. Rather than scattering proposed model corrections across many publications, it is critical to integrate these corrections in a central resource to ensure their availability to the scientific community in general, and to future

*To whom correspondence should be addressed. Tel: +1 650 859 5872; Fax: +1 650 859 3735; Email: keseler@ai.sri.com
Correspondence may also be addressed to Peter D. Karp. Tel: +1 650 859 4358; Fax: +1 650 859 3735; Email: pkarp@ai.sri.com

Table 1. EcoCyc content and *E. coli* gene product functions

Data type	Number
Genes	4499
Gene products covered by a mini-review	3706
Gene products with GO terms with EXP evidence	2462
Enzymes	1485
Metabolic reactions	1577
Compounds	2363
Transporters	264
Transport reactions	348
Transported substrates	254
Transcription factors	188
Regulatory interactions	5827
Transcription initiation	3207
Transcription attenuation	20
Regulation of translation	114
Enzyme modulation	2468
Other	18
Literature citations	23 909

modeling efforts in particular. Furthermore, general users of the database will benefit from the inclusion within the database of information required by the modeling effort, such as gene knockout phenotypes.

A second merit of our approach is that it will yield a more efficient and transparent modeling effort. Metabolic models require carefully curated lists of reactions, of chemical structures and of gene–protein–reaction relationships, and by directly leveraging the results of EcoCyc curation in a modeling effort, we can greatly reduce the amount of model-specific curation that is needed. Modeling efforts also require large amounts of data for evaluating model correctness, such as growth assays of the organism under many different nutrient and gene knockout conditions. Gathering, integrating and arbitrating among these data sets when they disagree can require substantial effort and can be carried out effectively within a MOD project. Modeling also becomes more efficient if successive phases of modeling build on the model corrections formulated in earlier phases, which is simplified if all model corrections are aggregated in a central database. Interpretation of model results and debugging of model errors will be accelerated by the ability to quickly access information about gene–protein–reaction relationships, regulatory information, genome arrangements and known literature about each gene. Interpretation of model results is speeded by computational tools such as the ability to visualize the hundreds of reaction flux rates predicted by a metabolic model onto a complete metabolic map diagram. The same visualization tools also make these models more transparent to the larger scientific community—rather than making model results available as large, cryptic spreadsheets and other data files, model results can be interpreted relative to web-accessible databases with powerful visualization tools.

Thus, driven in part by metabolic modeling efforts and in part by the utility of genome-scale data sets to the larger *E. coli* community, another new direction for EcoCyc is the integration of multiple large-scale growth and gene essentiality data sets into EcoCyc.

EcoCyc is part of the BioCyc collection of organism-specific pathway/genome databases at http://BioCyc.org (2). Among the nearly 2000 BioCyc databases are >130 databases for sequenced strains of *E. coli* and *Shigella*, including pathogenic, non-pathogenic, human microbiome and laboratory strains. These databases were automatically built using the Pathway Tools software and were not human-curated. Leveraging the large amount of experimental data collected in EcoCyc, we have begun to transfer functional annotations of gene products from *E. coli* K-12 MG1655 to their orthologs in the closely related K-12 strain W3110 and the *E. coli* B strain REL606. This allows us to focus manual curation efforts on the gene products and functions that differ between these strains.

UPDATE ON EcoCyc DATA

Update of transcriptional regulation data

Regulation of transcription initiation in EcoCyc has been kept up to date with the experimental literature. Table 2 summarizes the type and the number of regulation objects that are present in EcoCyc, as well as new objects added in the past 2 years. In addition, we have added missing evidence codes and literature citations to 210 promoters and to 85 regulatory interactions. All promoters and regulatory interactions now have at least one evidence code and one reference.

Allosteric regulation of RNA polymerase by guanosine tetraphosphate (ppGpp) and DksA

Regulation of transcription initiation goes beyond activator and repressor proteins that bind to the chromosome near promoter sequences. We have started curation of the alarmone guanosine tetraphosphate (ppGpp or 'magic spot') and the small protein DksA, both capable of binding RNA polymerase and thereby negatively regulating transcriptional activity of ribosomal RNA and transfer RNA genes in response to nutritional stress (3,4). ppGpp and DksA stimulate expression of proteins required for amino acid biosynthesis and transport (5–7), and some σ^E-dependent promoters (8) (Supplementary Figure S1). A total of 70 allosteric interactions have been curated; 29 are associated with ppGpp, 10 with DksA and 31 with both factors.

Improved annotation of transcription factor binding sites (TFBSs)

We continued updating and assigning the symmetry, length and consensus sequence for 130 transcription factors (TFs). As a consequence of this dedicated curation, we have relocated and reassigned TF binding sites (TFBSs) for 33 TFs and generated new regulatory interactions.

We used different strategies to identify the properties of the TFBSs, performed manual alignments of the regions upstream of the genes regulated, compared orthologous intergenic regions and used the information from other databases, such as PRODORIC (9), RegPrecise (10), Tractor_DB (11) and FITBAR (12). In all cases, we also

Table 2. Types and numbers of EcoCyc regulation objects

Data type	Total	New with high-confidence experimental evidence	New with computational or low-confidence experimental evidence
Transcription units	3473	19	48
Promoters	3766	53	1847
Terminators	251	0	12
TFs[a]	188	11	0
TFBSs	2701	183	144
Regulatory interactions	3207	69	412

[a]TFs include DNA-binding TFs, as well as RNA polymerase-binding regulators.

analyzed the available classical experimental evidence that corresponded to each regulatory interaction. All binding sites for a given TF were analyzed at the same time, using the biological knowledge of the mechanisms of action of TFs, preferential position and number of TFBSs, simple and complex regulation and families of regulators, among other aspects.

Improved position weight matrices (PWMs) and computational predictions of TFBSs

Computational predictions of TFBSs strongly rely on the quality of the position weight matrices (PWMs) that are used to scan regulatory regions, and on the threshold for selecting or rejecting TFBS predictions. To build a matrix, a minimum of four non-overlapping annotated binding sites is required for each TF. Medina-Rivera *et al.* (13) published a tool to assess the quality of matrices and define the appropriate score threshold. This tool has been used to evaluate and improve the matrices used to predict sites in *E. coli* K-12 regulatory regions. This evaluation, together with the continued detailed curation of TFBSs mentioned earlier and the increase in experimentally determined binding sites, has helped to increase the reliability of TFBS prediction. In 2010, a total of 11 522 binding sites were predicted for 71 TFs, whereas in 2012, we predicted fewer sites, 8718 improved predictions, for a larger set of 83 TFs. The improved PWMs were used to curate a set of regulatory interactions that had no binding site identified despite having experimental evidence that supported them. Our current manual curation of the predicted sites has identified TFBSs for 35 interactions.

Computationally predicted promoters

In addition to the curation of literature, we recently added 1852 computationally predicted promoters. The predicted potential promoters contain sequences similar to those recognized by six of the seven known sigma factors in *E. coli*: σ^{24}, σ^{28}, σ^{32}, σ^{38}, σ^{54} and σ^{70}. These predictions were generated by scanning 250-bp regions upstream of genes that lack reported promoters with PWMs for each sigma factor. PWMs and predictions for σ^{70} housekeeping promoters were generated as reported by Huerta and Collado-Vides in 2003 (14). An updated version of this strategy was used to generate PWMs and predictions for all other sigma factors, except for σ^{19}, as there is only one reported σ^{19}-dependent promoter.

Updates to the transport systems in EcoCyc

Curation of the *E. coli* transport systems in EcoCyc focuses foremost on the addition of new or significant functional characterizations reported in the literature. Since 2010, new transport functions have been assigned to nine previously uncharacterized membrane proteins (Supplementary Table S1). Motivated by the computational analyses described earlier, extensive literature searches resulted in the addition of transport reactions for a further 33 compounds in EcoCyc (Supplementary Table S2).

First, a long-running project to assess the dead-end metabolites identified within EcoCyc (15) continues to yield valuable information regarding the transport capabilities of *E. coli* K-12. A more detailed description of this analysis is the subject of a separate article.

Second, comparison of the experimental growth phenotypes recorded in EcoCyc with phenotypes predicted by computer modeling highlighted instances where transport reactions might be missing in EcoCyc. For example, *E. coli* K-12 is able to use pyruvate as a sole carbon and energy source (16), but no transport reaction for pyruvate was present in EcoCyc. A search of the literature revealed further information regarding the energetics of pyruvate transport in *E. coli* K-12 (17), and a reaction representing the transport of pyruvate across the inner membrane was thus added to EcoCyc, although the corresponding transporter is unknown. A total of 58 nutrient sources (carbon, nitrogen, sulfur or phosphorous) known to be capable of supporting growth were assessed in this way. Compounds for which transport reactions were added are shown in Supplementary Table S2. Literature references supporting the assertions of transport can be accessed via EcoCyc.

Review of transport protein nomenclature

EcoCyc has been curating transport proteins and transport reactions for many years, and over time, a variety of different curator approaches has resulted in transport protein nomenclature that was inconsistent and sometimes obscure. In many cases, the subunits of transport complexes were identified only by their gene name. Our aim in reviewing the nomenclature is to introduce a set of guidelines that will enable curators to apply consistent informative names to transport proteins, transport complexes and their subunits. In doing so, we have taken into account the prokaryotic protein naming guidelines developed by UniProt and also the International Union of Biochemistry and Molecular Biology (IUBMB)-approved classification system for membrane transport proteins known as the Transporter Classification system (18).

EcoCyc transport protein names are now indicative of substrate and transport energetics where possible. Gene names are not generally included, except in cases where more than one enzyme catalyses the same reaction. Transport class acronyms are retained if thought to be

widely recognizable (e.g. for ABC transporters and PTS permeases), but removed if less common. The individual subunits of transport complexes are named with the complex name followed by a specific subunit name. Table 3 lists examples of old and new transport protein names in EcoCyc that are illustrative of the improvements made. Approximately 450 individual transport proteins have been renamed in EcoCyc.

INTEGRATION OF PHENOTYPE DATA SETS

The full set of conditions that are suitable to sustain life for a bacterium is a fundamental collection of knowledge for that organism. Therefore, we have integrated data for 18 individual growth media and large-scale respiration measurements from five phenotype microarray (PM) experiments [(19,20,21); B. Bochner and X. Lei, personal communication; A. Mackie and I. Paulsen, personal commmunication]. Each PM experiment assays respiration (which is often treated as a proxy for growth) in one or more of a set of four 96-well plates containing a large set of standardized nutrient mixtures, and each well is counted as one growth observation. We integrated 1422 PM growth observations under aerobic conditions and 190 growth observations under anaerobic conditions. Many differences among these PM observations were detected and will be discussed in detail in a separate publication.

A summary of all growth conditions present in EcoCyc is available on the All Growth Media page, which can be retrieved through the command Search->Growth Media and then clicking the button 'All Growth Media for this Organism'. The first table in this web page lists individual growth media; subsequent tables list data for PMs (Figure 1). A button above the first table allows the user to select aerobic versus anaerobic conditions and wild type versus mutant strains. The PM tables are color coded to indicate the degree of growth observed. When multiple observations are available for a given cell, the color of the cell is determined as follows. If all observations agree (e.g. all observations indicate growth), then the color of the cell indicates that growth level (e.g. growth). If the observations differ, but a curator has arbitrated among them and assigned a consensus result (e.g. for citrate, no growth was observed for most observations, and no growth was also observed by low-throughput experiments in the literature), the overall color of the cell reflects that consensus (no growth), but a small grid within the cell shows the individual observations (move the mouse over an element of that grid for a citation to each experiment). If the observations differ, and a curator has not arbitrated among the differences, then the overall color of the cell indicates that the observations are inconsistent, and the small grid within the cell shows the individual observations.

Clicking on a cell within the All Growth Media page produces a page describing all its chemical components and listing all growth observations available for that growth medium across all available conditions and mutants. For example, the medium 'MOPS medium with 0.4% glucose' lists growth observations for 4214 single-gene knockouts of *E. coli*.

Gene knockout data

Gene essentiality information is useful for predicting antibiotic targets for pathogenic bacteria and for guiding the design of minimal genomes. It provides clues regarding the functions of genes of unknown function. Additionally, it is useful for validating genome-scale metabolic flux models because those models can simulate the effects of knockouts; model results are compared with the experimental data to assess model accuracy. We have loaded five high-throughput gene knockout data sets into EcoCyc (22–26) that include >13 000 individual gene-knockout growth observations. Each growth observation is tied to the growth medium in which the observation was made, as the notion of gene essentiality depends strongly on the conditions under which essentiality is assessed. Gene knockout phenotypes are shown both on the growth medium page and in a table within the gene page (Figure 2).

BEYOND *E. coli* K-12 MG1655

The *E. coli* strain K-12 MG1655 was chosen for genome sequencing because it had undergone comparatively minimal genetic manipulation since its isolation; the completed sequence was published in 1997 (27) and updated in 2006 (28,29). Since then, several other commonly used laboratory strains, as well as many pathogenic and commensal strains of *E. coli*, have been fully sequenced. Because of the large number of genome sequences, manual curation of even a small subset of the resulting databases is neither feasible nor efficient.

Table 3. Examples of renamed transport proteins in EcoCyc

Former name	Revised name
YdeA MFS transporter	arabinose efflux transporter
rhamnose RhaT transporter	rhamnose/lyxose:H^+ symporter
GabP APC transporter	4-aminobutyrate:H^+ symporter
MglB	galactose ABC transporter—periplasmic binding protein
MglC	galactose ABC transporter—membrane subunit
MglA	galactose ABC transporter—ATP-binding subunit
CorA magnesium ion MIT transporter	Ni^{2+}/Co^{2+}/Mg^{2+} transporter
MalX	maltose/glucose PTS permease—MalX subunit
EmrE SMR transporter	multidrug efflux transporter EmrE

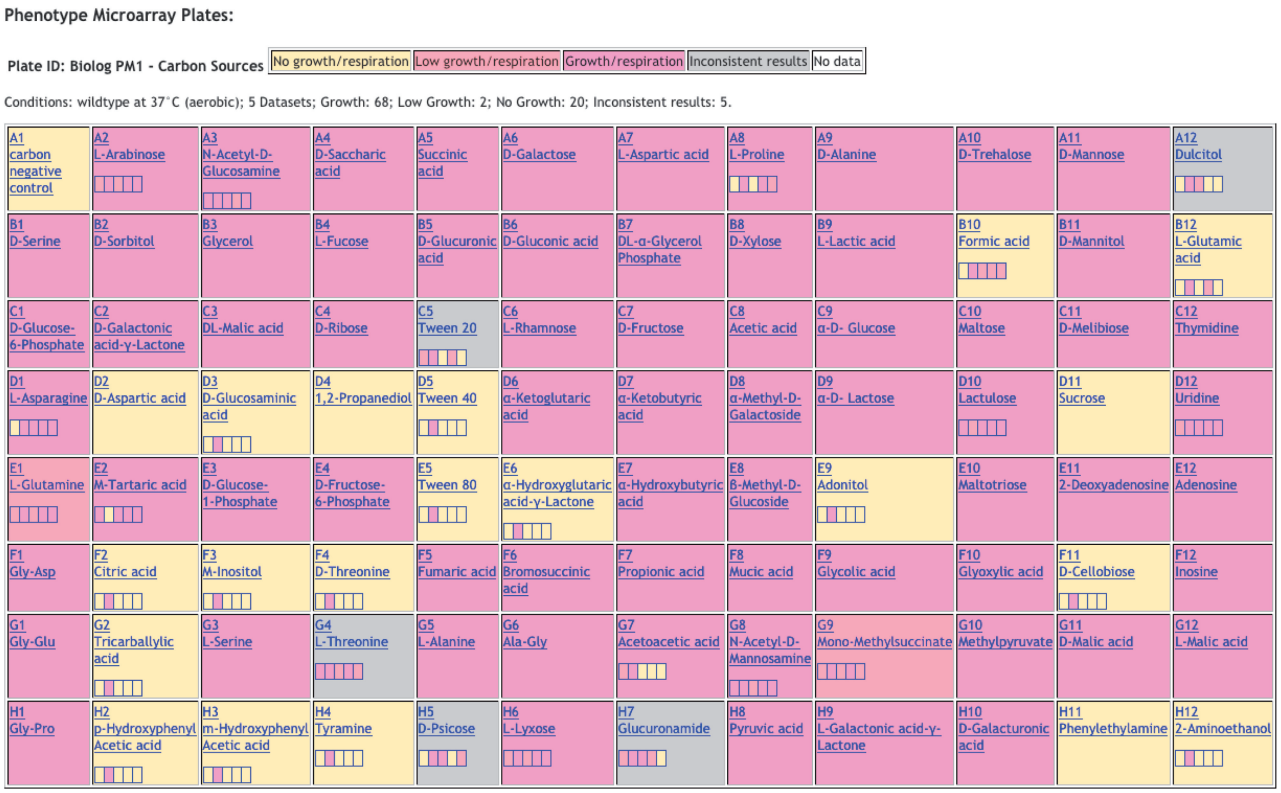


Figure 1. Biolog PM1 plate depicting *E. coli* carbon source utilization results from five different experiments under aerobic growth conditions.

To address this problem, the Pathway Tools software now includes automated and manual tools for curators to transfer annotations from a well-curated 'master' MOD to orthologs in databases of its less-well curated relatives.

To limit the likelihood of inappropriate transfer of annotations, the criteria used by the automated tool are strict. Candidate gene pairs are identified on the basis of sequence orthology, defined as the best bidirectional BLAST hit. In addition, cutoffs for alignment quality (BLAST *P*-value of 10^{-10}), alignment length and synteny are enforced, and the presence of existing annotations that may conflict with transferred annotations is taken into account. Functions of individual orthologs that do not meet all of these criteria, but should nevertheless be transferred, can be propagated by a curator. The values copied from genes/proteins in the 'master' database include the gene and gene product names and synonyms, heteromultimeric complexes, reactions catalyzed by proteins and complexes and GO terms with experimental evidence codes.

We have initially transferred annotations from EcoCyc to orthologous genes in the BioCyc.org databases for the K-12 strain W3110 and the B strain REL606. Manual updates to orthologs in both databases are under way. Well-known differences between the metabolic capabilities of the K-12 and B strains will be captured in our current curation effort.

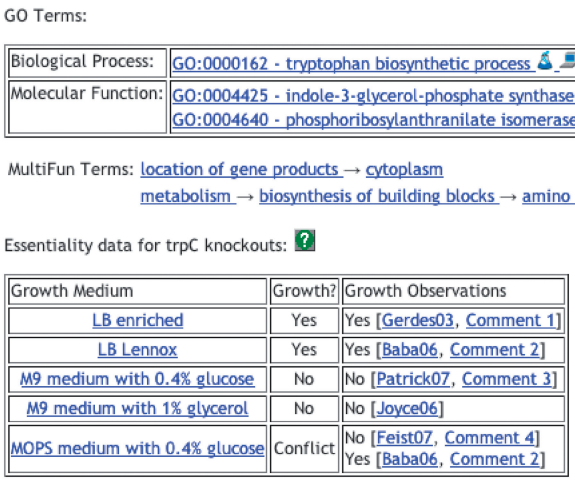
EcoCyc METABOLIC FLUX MODEL

The MetaFlux software generates steady-state metabolic flux models from pathway/genome databases (1). This approach ensures that updates to the database are automatically reflected in the generated model. We have generated (1) a FBA model for EcoCyc that can be executed using MetaFlux as part of the downloadable software/database bundle that includes Pathway Tools and EcoCyc; the model is also available as an SBML file within the EcoCyc downloadable files (http://biocyc.org/download.shtml).

The EcoCyc FBA model comprises 1888 total reactions; the model produces 58 biomass metabolites with 370 reactions carrying non-zero flux, from a minimal medium that includes glucose and ammonium. We assessed the accuracy of the model against the growth observations and gene knockout data in EcoCyc. The model predicted growth versus no growth correctly for 72.6% of 383 growth conditions in EcoCyc. The model predicted growth versus no growth for the 4207 single-gene knockouts in (21) with 91.2% accuracy.

WEB INTERFACE UPDATES: WEB GROUPS

Web Groups are a new aspect of the EcoCyc/BioCyc web site that allow users to create, store, analyze and display groups of genes, metabolites, pathways and other entities within EcoCyc. Groups can also be shared with specific colleagues or made fully public. Although a full description of Web Groups is beyond the scope of this article, we provide here a sample use scenario for Web Groups. We will create a Web Group containing a set of *E. coli* genes of interest (e.g. from a gene expression experiment or from


GO Terms:

Biological Process:	GO:0000162 - tryptophan biosynthetic process [GOA00, Smith67b]
Molecular Function:	GO:0004425 - indole-3-glycerol-phosphate synthase activity [GOA01a, GOA01, Creighton66] GO:0004640 - phosphoribosylanthranilate isomerase activity [GOA01a, GOA01, Hommel95, Smith67b]

MultiFun Terms: location of gene products → cytoplasm
metabolism → biosynthesis of building blocks → amino acids → tryptophan

Essentiality data for trpC knockouts:

Growth Medium	Growth?	Growth Observations
LB enriched	Yes	Yes [Gerdes03, Comment 1]
LB Lennox	Yes	Yes [Baba06, Comment 2]
M9 medium with 0.4% glucose	No	No [Patrick07, Comment 3]
M9 medium with 1% glycerol	No	No [Joyce06]
MOPS medium with 0.4% glucose	Conflict	No [Feist07, Comment 4] Yes [Baba06, Comment 2]

Figure 2. The section of an EcoCyc gene page that provides gene essentiality information.

some other type of experiment), and use several tools to determine commonalities among that set of genes.

To begin using Web Groups, a user must first create an account (groups are stored within a user's account) and start a Web Groups session (command Tools->Groups). A Groups command menu then becomes available in the menu bar.

Web Groups can be created in several ways, e.g. by uploading the gene list from a file, or from a list of search results. Most EcoCyc object pages also allow you to add an object (e.g. a metabolite) to an existing group. Once a new group is created, a single column of information will be shown, namely the name of each gene. Additional properties of these EcoCyc gene objects can be selected.

One way to see whether the genes in this group are found in a common set of metabolic pathways is to create a new column in which we transform each gene to the pathways its product is present in by selecting the 'Pathways of gene' transform. That column can be converted to a new group by clicking on the '+' in the column heading. Many other transformations are available, for example, a gene group can be transformed to a list of all genes that are regulated by genes in the group, and to a list of all orthologous genes in another organism. The list of transformations available depends on the type of objects within the current group.

Another way to investigate pathway relationships within a gene group is to highlight those genes on the EcoCyc metabolic map diagram, which we call the Cellular Overview.

A final way of investigating shared relationships among these genes is by using enrichment analysis, which is a statistical technique for determining whether a set of entities (such as our gene list) is statistically over-represented for members of other known sets. For example, does our gene set contain more genes from a given metabolic pathway than we expect by chance? The 'Enrichments' menu allows you to apply several statistical tests to a group; the exact tests available depend on the type of objects within the group. Currently, enrichment analyses are available for gene groups and metabolite groups. In addition to testing a gene group for pathway enrichment, tests for enrichment for GO terms are available, as is a test for whether the genes in a group share regulators in common more frequently than would be expected by chance. Finally, you can perform all of these tests at once and see the results sorted by *P*-value.

More information about Web Groups is available through a BioCyc Webinar (http://www.biocyc.org/webinar.shtml) and from the BioCyc Web site User's Guide (http://www.biocyc.org/BioCycUserGuide.shtml). For example, groups can be exported to files. We expect that in late 2012, set operations will be available for groups, and it will be possible to manipulate sequences using groups.

USING EcoCyc AS A TEACHING RESOURCE

With the goal of assisting undergraduate student learning of microbiology principles in large classroom settings, we are exploiting EcoCyc for college-level instruction. This approach uses web-based tutorials to orient the student in accessing and using EcoCyc. Student learning modules introduce basic microbiological principles, and a set of complementary student exercises is designed to reinforce topics covered in formal classroom lectures. We reason that web-based exercises can deepen student understanding of basic microbiology concepts and improve overall class performance. The web-based educational approach also allows for independent and self-paced learning while increasing the depth of inquiry and study, which is not easily accomplished in large classroom settings.

We authored additional EcoCyc-based educational modules and evaluated their effectiveness in an introductory-level microbiology lecture course of 255 students at UCLA in the spring of 2012. These modules describe basic principles of *E. coli* nutrient uptake, energy generation by aerobic and anaerobic respiration, substrate-level phosphorylation, fermentation, genome organization, gene regulation and genome/organism comparison. The materials are accessible at the newly created *E. coli* student portal web site at http://ecolistudentportal.org.

How the learning materials were implemented

Following the introduction of a basic microbiological principle/process in class lecture, the instructor assigned a web-based task to each student. After performing a web-based research inquiry, each student answered a set of questions to demonstrate mastery of the assigned topic—for example, to identify genes/gene product relationships for specific membrane transport systems, the enzyme system affinities, and specificities for substrate(s). The student also provided a brief statement explaining the rationale and approach used. Exercises were graded for accuracy and completeness. The major goals of each task in this project were to have each student demonstrate understanding of the class-introduced concepts, master use of a state-of-the-art microbial MOD and stimulate inquiry-based learning of a topic beyond the class lecture.

How the materials were evaluated

A student survey was conducted at the end of the course using the tools at http://www.salgsite.org/. Questions were designed to measure student perceptions of learning gains made as the result of the EcoCyc-based exercises. Student response rate was >91%, and the mean response and confidence intervals were determined for all replies.

Outcomes

For student perceptions, 97% of the students successfully completed all assigned web tasks that comprised 10% of the final course grade. For the assigned exercises, the mean student scores ranged from 91 to 96%, with a low score of 71% to a high score of 100%. This was the first class exposure to a web-based MOD learning experience, and a majority of the students were excited about using a research grade tool to access and analyze data regarding *E. coli* biology. Responses to the student exit poll revealed that the goals of the exercises were generally clear, were relevant to class material and reinforced learning of general microbiological principles. Our instructional goal to have every student demonstrate proficiency in self-directed inquiry using the EcoCyc database was nearly achieved.

Future directions

We plan to author additional *E. coli* learning modules and complementary exercises in other areas of basic microbiology using the EcoCyc database as a web-based resource. Outreach to other institutions is planned to facilitate sharing of course materials and to assist in classroom implementation of these materials.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1 and 2 and Supplementary Figure 1.

ACKNOWLEDGEMENTS

The authors thank Sydney Kustu for first suggesting the idea of 'navigation by groups'. The content of this article is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences or the National Institutes of Health.

FUNDING

National Institute of General Medical Sciences of the National Institutes of Health [U24GM077678, GM088849 to P.D.K., GM071962 to J.C.-V.]. Funding for open access charge: NIH [U24GM077678].

Conflict of interest statement. SRI authors benefit from a commercial licensing program for Pathway Tools.

REFERENCES

1. Latendresse,M., Krummenacker,M., Trupp,M. and Karp,P.D. (2012) Construction and completion of flux balance models from pathway databases. *Bioinformatics*, **28**, 388–396.
2. Caspi,R., Altman,T., Dreher,K., Fulcher,C.A., Subhraveti,P., Keseler,I.M., Kothari,A., Krummenacker,M., Latendresse,M., Mueller,L.A. *et al*. (2012) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res*., **40**, D742–D753.
3. Paul,B.J., Barker,M.M., Ross,W., Schneider,D.A., Webb,C., Foster,J.W. and Gourse,R.L. (2004) DksA: a critical component of the transcription initiation machinery that potentiates the regulation of rRNA promoters by ppGpp and the initiating NTP. *Cell*, **118**, 311–322.
4. Lemke,J.J., Sanchez-Vazquez,P., Burgos,H.L., Hedberg,G., Ross,W. and Gourse,R.L. (2011) Direct regulation of *Escherichia coli* ribosomal protein promoters by the transcription factors ppGpp and DksA. *Proc. Natl Acad. Sci. USA*, **108**, 5712–5717.
5. Barker,M.M., Gaal,T., Josaitis,C.A. and Gourse,R.L. (2001) Mechanism of regulation of transcription initiation by ppGpp. I. Effects of ppGpp on transcription initiation in vivo and in vitro. *J. Mol. Biol*., **305**, 673–688.
6. Srivatsan,A. and Wang,J.D. (2008) Control of bacterial transcription, translation and replication by (p)ppGpp. *Curr. Opin. Microbiol*., **11**, 100–105.
7. Paul,B.J., Berkmen,M.B. and Gourse,R.L. (2005) DksA potentiates direct activation of amino acid promoters by ppGpp. *Proc. Natl Acad. Sci. USA*, **102**, 7823–7828.
8. Costanzo,A., Nicoloff,H., Barchinger,S.E., Banta,A.B., Gourse,R.L. and Ades,S.E. (2008) ppGpp and DksA likely regulate the activity of the extracytoplasmic stress factor sigmaE in *Escherichia coli* by both direct and indirect mechanisms. *Mol. Microbiol*., **67**, 619–632.
9. Grote,A., Klein,J., Retter,I., Haddad,I., Behling,S., Bunk,B., Biegler,I., Yarmolinetz,S., Jahn,D. and Munch,R. (2009) PRODORIC (release 2009): a database and tool platform for the analysis of gene regulation in prokaryotes. *Nucleic Acids Res*., **37**, D61–D65.
10. Novichkov,P.S., Brettin,T.S., Novichkova,E.S., Dehal,P.S., Arkin,A.P., Dubchak,I. and Rodionov,D.A. (2012) RegPrecise web services interface: programmatic access to the transcriptional regulatory interactions in bacteria reconstructed by comparative genomics. *Nucleic Acids Res*., **40**, W604–W608.
11. Perez,A.G., Angarica,V.E., Vasconcelos,A.T. and Collado-Vides,J. (2007) Tractor_DB (version 2.0): a database of regulatory interactions in gamma-proteobacterial genomes. *Nucleic Acids Res*., **35**, D132–D136.
12. Oberto,J. (2010) FITBAR: a web tool for the robust prediction of prokaryotic regulons. *BMC Bioinformatics*, **11**, 554.
13. Medina-Rivera,A., Abreu-Goodger,C., Thomas-Chollier,M., Salgado,H., Collado-Vides,J. and van Helden,J. (2011) Theoretical and empirical quality assessment of transcription factor-binding motifs. *Nucleic Acids Res*., **39**, 808–824.

14. Huerta,A.M. and Collado-Vides,J. (2003) Sigma70 promoters in *Escherichia coli*: specific transcription in dense regions of overlapping promoter-like signals. *J. Mol. Biol.*, **333**, 261–278.
15. Karp,P.D., Keseler,I.M., Shearer,A., Latendresse,M., Krummenacker,M., Paley,S.M., Paulsen,I., Collado-Vides,J., Gama-Castro,S., Peralta-Gil,M. *et al.* (2007) Multidimensional annotation of the *Escherichia coli* K-12 genome. *Nucleic Acids Res.*, **35**, 7577–7590.
16. Bochner,B.R., Gadzinski,P. and Panomitros,E. (2001) Phenotype microarrays for high-throughput phenotypic testing and assay of gene function. *Genome Res.*, **11**, 1246–1255.
17. Lang,V.J., Leystra-Lantz,C. and Cook,R.A. (1987) Characterization of the specific pyruvate transport system in *Escherichia coli* K-12. *J. Bacteriol.*, **169**, 380–385.
18. Saier,M.H. Jr, Tran,C.V. and Barabote,R.D. (2006) TCDB: the transporter classification database for membrane transport protein analyses and information. *Nucleic Acids Res.*, **34**, D181–D186.
19. AbuOun,M., Suthers,P.F., Jones,G.I., Carter,B.R., Saunders,M.P., Maranas,C.D., Woodward,M.J. and Anjum,M.F. (2009) Genome scale reconstruction of a Salmonella metabolic model: comparison of similarity and differences with a commensal *Escherichia coli* strain. *J. Biol. Chem.*, **284**, 29480–29488.
20. Baumler,D.J., Peplinski,R.G., Reed,J.L., Glasner,J.D. and Perna,N.T. (2011) The evolution of metabolic networks of *E. coli*. *BMC Syst. Biol.*, **5**, 182.
21. Yoon,S.H., Han,M.J., Jeong,H., Lee,C.H., Xia,X.X., Lee,D.H., Shim,J.H., Lee,S.Y., Oh,T.K. and Kim,J.F. (2012) Comparative multi-omics systems analysis of Escherichia coli strains B and K-12. *Genome Biol*, **13**, R37.
22. Baba,T., Ara,T., Hasegawa,M., Takai,Y., Okumura,Y., Baba,M., Datsenko,K.A., Tomita,M., Wanner,B.L. and Mori,H. (2006) Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.*, **2**, 2006.0008.
23. Joyce,A.R., Reed,J.L., White,A., Edwards,R., Osterman,A., Baba,T., Mori,H., Lesely,S.A., Palsson,B.O. and Agarwalla,S. (2006) Experimental and computational assessment of conditionally essential genes in *Escherichia coli*. *J. Bacteriol.*, **188**, 8259–8271.
24. Patrick,W.M., Quandt,E.M., Swartzlander,D.B. and Matsumura,I. (2007) Multicopy suppression underpins metabolic evolvability. *Mol. Biol. Evol.*, **24**, 2716–2722.
25. Gerdes,S.Y., Scholle,M.D., Campbell,J.W., Balazsi,G., Ravasz,E., Daugherty,M.D., Somera,A.L., Kyrpides,N.C., Anderson,I., Gelfand,M.S. *et al.* (2003) Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *J. Bacteriol.*, **185**, 5673–5684.
26. Feist,A.M., Henry,C.S., Reed,J.L., Krummenacker,M., Joyce,A.R., Karp,P.D., Broadbelt,L.J., Hatzimanikatis,V. and Palsson,B.O. (2007) A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.*, **3**, 121.
27. Blattner,F.R., Plunkett,G. 3rd, Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F. *et al.* (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1462.
28. Hayashi,K., Morooka,N., Yamamoto,Y., Fujita,K., Isono,K., Choi,S., Ohtsubo,E., Baba,T., Wanner,B.L., Mori,H. *et al.* (2006) Highly accurate genome sequences of *Escherichia coli* K-12 strains MG1655 and W3110. *Mol. Syst. Biol.*, **2**, 2006.0007.
29. Riley,M., Abe,T., Arnaud,M.B., Berlyn,M.K., Blattner,F.R., Chaudhuri,R.R., Glasner,J.D., Horiuchi,T., Keseler,I.M., Kosuge,T. *et al.* (2006) *Escherichia coli* K-12: a cooperatively developed annotation snapshot–2005. *Nucleic Acids Res.*, **34**, 1–9.